Combining Mathematical and Statistical Modeling to Simulate Time Course Bulk and Single Cell Gene Expression Data

Thomas D Sherman[1], Luciane T Kagohara[1], Raymon Cao[1], Raymond Cheng[2], Matthew Satriano[3], Michael Considine[1], Gabriel Krigsfeld[1], Ruchira Ranaweera[4], Yong Tang[5], Sandra A Jablonski[6], Genevieve Stein-O'Brien[1,7], Daria A Gaykalova[8], Louis M Weiner[6], Christine H Chung[4], and Elana J Fertig[1]

[1] Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD USA, [2] Science, Math and Computer Science Magnet Program, Poolesville High School, Poolesville, MD USA , [3] Department of Mathematics, University of Waterloo, Waterloo, Ontario, Canada, [4] Moffitt Cancer Center, Tampa, FL, USA, [5] Salubris Biotherapeutics, Inc, Gaithersburg, MD, USA, [6] Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC USA, [7] Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD USA, [8] Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD USA

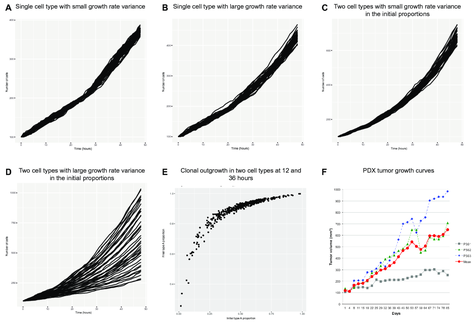
Abstract

Motivation: Bioinformatics techniques to analyze time course bulk and single cell omics data are advancing. The absence of a known ground truth of the dynamics of molecular changes challenges benchmarking their performance on real data. Realistic simulated time-course datasets are essential to assess the performance of time course bioinformatics algorithms.

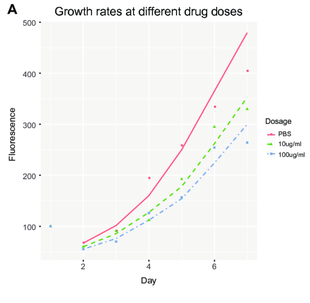
Results: We develop an R/Bioconductor package *CancerInSilico* to simulate bulk and single cell transcriptional data from a known ground truth obtained from mathematical models of cellular systems. This package contains a general infrastructure for running cell-based mathematical models and pipelining the results into a pathway simulation which forms the basis for generating mean expression data. The mean expression data is run through an appropriate error model to generate the final simulated data.

Cell-Based Model

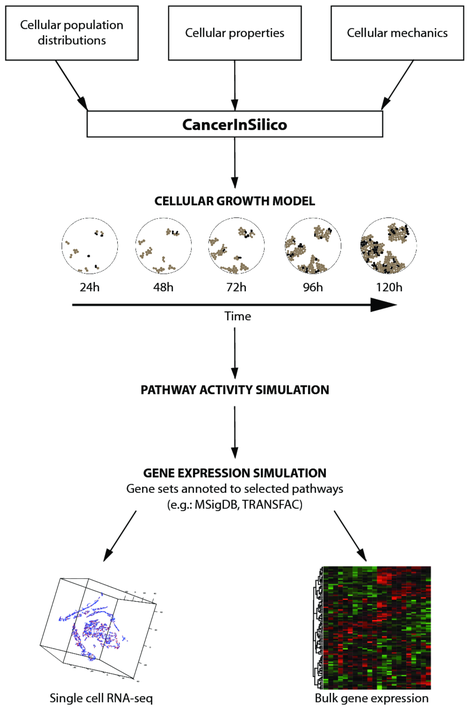
While the package infrastructure allows for multiple cell-based models, the current version comes with just a single model implemented. This model is an off-lattice, stochastic cell-based model adapted from the work of Drasdo and Höhme (Drasdo and Höhme, 2003). It is a simple, two-dimensional model of cell growth that captures properties such as cell-cell adhesion and limited cell deformability. This model is extended to include multiple cell types and their response to drug therapies.



Comparison of simulated cell growth with multiple cell types and in vivo tumor growth of a patient derived xenograft model. *(a) Cellular growth over time in a set of 200 simulations containing a single cell type. (b) As for (a) with greater growth rate standard deviation. (c) Cellular growth over time in a simulation containing a two cell types with large difference in mean growth rates. Plots represent 40 simulations with initial proportion of cells of type 1 selected from a uniform between 0.45 and 0.55. (d) As for (c) with initial proportion of cells of type 1 selected from a uniform between 0 and 1. (e) Proportion of cells of type 1 in the initial population vs final population. (f) Tumor volume of triplicates for a single patient derived xenograft model.*



Comparison of simulated and in vitro cellular growth with and without cetuximab treatment. *Simulated growth curves (lines) and in vitro growth (dots) for BxPC-3 cells treated with 10 µg/ml or 100 µg/ml of cetuximab and PBS controls.*



Pathway Simulation

The connection between the cell model and the gene expression simulation happens through user defined pathways that are associated with gene expression changes in a corresponding set of genes annotated to that pathway. We define an intermediate variable (*P*) that is a continuous value between zero and one that records how active each biological pathway is within each modeled cell. This measure of pathway activity is used to scale the mean expression value within a range observed in real data.

$$G_{mean} = G_{min} + P \times (G_{max} - G_{min})$$

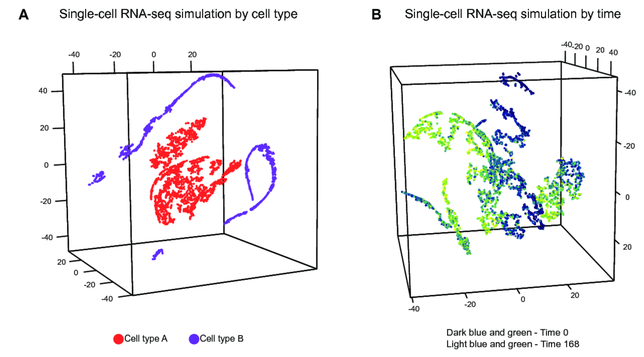
In this equation, G_{min} and G_{max} are estimated from observed data. For any particular gene, the mean expression is found for all simulated pathways the gene belongs to and the maximum of these values is used as the simulated mean value. For this study, we model four pathways related to the phase transition from G to S and G to M, growth factor signaling, and contact inhibition. For the G to M and G to S pathways, P is either 0 or 1 at the current time point depending on if the cell transitioned between phases within a specified time window. Activity in the growth factor signaling pathway is assumed to be proportional to the expected cell cycle length of each cell. The contact inhibition activity is defined by the "local density" of the cell, which is the proportion of area within a small radius of the reference cell that is occupied by other cells.

Simulating Gene Expression Data

Once the mean expression matrix is generated, additional genes unaffected by any pathway are also simulated with a pre-specified distribution. Technical error is then simulated in all genes appropriate to the measurement platform. A normal error model is used to simulate log transformed microarray data and a negative binomial error model adapted from the code for LIMMA voom (Law *et al.*, 2014) is used to simulate bulk RNA-sequencing data. Technical noise for simulated single cell RNA-sequencing data is generated using the error and drop out models from Splatter (Zappia *et al.*, 2017).

Analysis of Simulated Data

T-SNE of simulated time course single cell RNA-sequencing data for a population with cells of types A and B. *Points colored by (a) cell type, (b) time – both plots generated with the R package Rtsne.*



We apply this technique to simulate single cell RNA-sequencing data from a simulation of a population of cells equally distributed between two cell types. Each cell type has an associated pathway with a gene set that corresponds to cellular identity – i.e the pathway activity is either 0 or 1 depending on the cell type. In this analysis we observe strong separation between cell types and time, most likely due to the effects of contact inhibition being strongly correlated with time.

Acknowledgements

We thank Ludmila V Danilova, Alexander V Favorov, Emily Flam, Dylan Kelley, Feilim Mac Gabhann, Cristian Tomasetti, and members of NewPISlack for critical comments and feedback on this project. This project has been made possible in part by grant number 2018-183444 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

References

Drasdo,D. and Höhme,S. (2003) Individual-based approaches to birth and death in avascular tumors. *Math. Comput. Model.*, **37**, 1163–1175.

Law,C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.

Zappia,L. *et al.* (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**